Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings 'corrected' for frequency

Marc Brysbaert

Ghent University, Belgium

Key words: age of acquisition, word recognition, word frequency

To be published in QJEP

Address: Marc Brysbaert
Department of Experimental Psychology
Ghent University
Henri Dunantlaan 2
B-9000 Gent
Belgium
Tel. +32 9 264 94 25
Fax. +32 9 264 64 96
E-mal: marc.brysbaert@ugent.be

Abstract

From the very first studies researchers on the effect of age-of-acquisition (AoA) on word processing have validated their AoA ratings by correlating them with other, more objective indices of the age at which children know words. Still, the ratings have been questioned and alternative measures have been proposed. Two of these are differences in word frequency between language directed at young children and language directed at older children (frequency trajectory) and AoA ratings corrected for word frequency. Surprisingly, the criterion validity of these alternative measures has never been established, partly because one of the validation criteria (the age at which children are able to name pictures) has been questioned. In the present study, four databases are used that aimed to establish the order of English word acquisition, going from the very first words learned to words taught in secondary education. The criteria for word knowledge included word production, multiple choice questions about the meaning of the words, and teacher judgments about when words should be taught in the school curriculum. For all databases, the frequency trajectory correlated substantially less with the criterion than AoA ratings. For all but one, the same was true for AoA ratings 'corrected' for other variables. On the basis of these findings, researchers should be cautious interpreting null effects with the 'improved' variables as evidence against a genuine AoA effect.

Regression analysis of large samples of word processing times indicates that four variables account for most of the variance in reaction times (RTs). They are word frequency, word length, similarity to other words, and age of acquisition (Brysbaert, Stevens, Mandera, & Keuleers, 2016).[1] Words are easier to process (at least in a lexical decision task), when they have a high frequency, are short, resemble many other words, and have been acquired early. Of those variables, age of acquisition (AoA) is the most contested, because it is based on ratings provided by adults (mostly undergraduate students). The problematic issue is that subjective ratings may incorporate more information than pure AoA. For instance, participants may erroneously think they acquired high-frequency words earlier than low-frequency words. They also tend to underestimate the number of words learned before the age of four and after the age of 15. And it can be questioned whether participants have any recollection of when they learned words at a preschool age.

Authors who believe in the impact of AoA on word learning, have provided three arguments why AoA (as measured with ratings) has a genuine impact. The first is that AoA ratings correlate well with more objective measures of AoA. In the very first paper on the AoA effect in healthy participants, Carroll and White (1973) not only used AoA ratings for the names of 94 pictures, but also the grade at which the words were expected to be taught. The latter were based on the work of Edgar Dale and colleagues who spent nearly a lifetime trying to establish at which age words should be taught. These authors did so by testing word knowledge in large samples of children from different classes. One of the formats they used was a multiple choice test with three alternatives (see below). Carroll and White (1973) reported a correlation of .85 between the ratings and the grades. A similar procedure was followed by Gilhooly and Gilhooly (1980, Experiment 1). They used 53 words from the Mill Hill vocabulary test, which contains information about the frequency with which each word is known by children under 11 years of age. The correlation between the difficulty rank position in the Mill Hill test and the AoA rating was .93. In a second experiment, Gilhooly and

[1] To these variables, Brysbaert et al. (2016) added a fifth: word prevalence (how many people know the word in a vocabulary test). However, the impact of this variable so far has only been attested in the Dutch language.

Gilhooly (1980) presented 48 words with various AoA ratings to children from primary school and the first year of secondary school. The children were asked to say what each word meant. The objective AoA was determined as the age at which 50 percent of the children were able to give an acceptable meaning of the word. The correlation between rated AoA and objective AoA was .84. Morrison, Chappell, and Ellis (1997) tested even younger children (starting from 2:6 years up to 10:11 years), to whom they presented 297 pictures that had to be named. AoA was defined as the age at which 75% of the children could name the picture. The correlation between rated AoA and objective AoA was .75. Numerous other studies in other languages have confirmed these findings. For instance, De Moor, Ghyselinck, and Brysbaert (2000, Experiment 1) asked 80 children of 5-6 years to define 249 Dutch words. The correlation between the percentage of children that knew the answer (coded by three independent judges on the basis of tape recordings) and rated AoA was .75.

The second path taken by researchers to show a genuine effect of AoA on word processing has been to design learning studies, in which more control can be exerted about the order and the frequency with which stimuli are taught. For instance, Tamminen and Gaskell (2008) had participants learn new pseudowords in different sessions. When the participants had to name the pseudowords a few weeks later, they showed an order of acquisition effect: The first learned words were named faster than the last learned words. Catling, Dent, Preece, and Johnston (2013) taught participants the names of personlike figures (Greebles). The training regime was such that the cumulative presentation frequency remained constant. Still, the authors observed a difference in naming latency between the first presented figures and the later presented figures. Measuring eye movements, Joseph, Wonnacott, Forbes, and Nation (2015) observed a similar order of acquisition effect for newly learned words in text reading

Finally, the third path taken by researchers to make a genuine AoA effect acceptable was to show how such an effect could arise from computational models. Ellis and Lambon Ralph (2000) showed that a connectionist network resulted in an order of acquisition effect when the relationship between

input and output was arbitrary and learning was interleaved (i.e., the first acquired input kept on being presented while the later acquired input was taught to the model). The order of acquisition effect remained even when the cumulative frequency was controlled for. Monaghan and Ellis (2010) later showed that naming English words contains enough arbitrariness in the letter-sound correspondences to show an AoA effect in word naming under a realistic training regime.

Despite the impressive range of evidence for a genuine AoA effect in word processing, arguments against it keep on being published. They are based on the fact that under some conditions the AoA effect is not observed in connectionist models (see Monaghan & Ellis, 2010, for a review of these conditions) and on the fact that the AoA effect in word recognition is not observed when 'better', statistical measures are used than AoA ratings. With respect to the latter, it is important to notice that the improved measures are not based on the objective, performance-based, estimates described above, which are available for a limited number of words only (though see below). Instead, they consist of two derived measures: the frequency trajectory measure and AoA ratings corrected for frequency (and other variables).

The frequency trajectory variable was first proposed by Zevin and Seidenberg (2002), who particularly questioned the AoA effect in skilled word naming. According to these authors, the correspondence between word orthography and phonology is not arbitrary enough for an AoA effect to emerge. They argued that the objective age of acquisition norms obtained by Morrison et al. (1997) could be explained by differences in childhood frequency and were no longer expected to affect adult performance because the order of acquisition is likely to be wiped out at that age as a result of the massive cumulative frequency with which words have been encountered since childhood. Zevin and Seidenberg (2004) also argued that picture naming in young children was not a good task to measure AoA, because pictures that are difficult to name in adulthood (for whatever reason) will also be difficult to name in childhood. So, the AoA differences observed by Morrison et al. could be an artefact of the effort to process the picture or to say the name.

Zevin and Seidenberg (2004) suggested that a better way to measure of AoA was to compare the frequencies of words at different ages. Early acquired words could be defined as words that are more frequent in childhood than later; late acquired words have a lower frequency in childhood than later. The authors tested the new frequency trajectory measure for 328 words and showed that it correlated .54 with rated AoA, but was unrelated to word frequency (unlike the AoA ratings, which correlated -.69 with word frequency). Next, they showed that frequency trajectory did not affect word naming times.

The idea of frequency trajectory was taken up by Bonin and colleagues. In a series of publications they gradually shifted from a belief in a genuine AoA effect to a rejection of the AoA effect on the basis of the frequency trajectory. First, Bonin, Barry, Meot, and Chalard (2004) showed a significant effect of frequency trajectory on word naming latencies in French, when word frequency was controlled for. Second, Bonin, Meot, Mermillod, Ferrand, and Barry (2009) argued that frequency trajectory should be used instead of rated AoA, even though it led to less variance explained. Finally, Lété and Bonin (2013) ran a series of studies which led them to conclude that "The findings firmly establish that in alphabetic languages such as French, age-limited learning effects do not surface readily in word recognition. In contrast, the total exposure to words across the lifetime is a strong determinant of word recognition speed." (p. 973). This conclusion was based on a series of 6 experiments in which words with different frequency trajectories were compared. The words with higher frequencies in the first grades of primary school than in the last grades resulted in significant effects on AoA ratings, but not on word recognition tasks, such as word naming, lexical decision, semantic decision, or progressive demasking (see also Bonin, Lété, Méot, Roux, & Ferrand, in press).

A different attack on AoA ratings was raised by Baayen and colleagues (e.g., Baayen, Milin, & Ramscar, in press). They also took issue with the high correlation between AoA ratings and word frequency, and argued that if one is interested in the effects of AoA, one should use AoA ratings

corrected for word frequency. The easiest way to do this is to use the residuals after AoA ratings are predicted on the basis of word frequency. Another method is to run a principal components analysis on word frequency and AoA, in which the first component extracts the information shared by word frequency and AoA, whereas the second, orthogonal component represents the difference between word frequency and AoA. Baayen et al. (in press) even went further and claimed that nearly all variance in AoA ratings could be explained on the basis of word frequency, word length, word valence (negative or positive affect), word arousal, number of similar words (N count), number of meanings of the word, and principal components related to the meaning of the words. Given this state of affairs, Baayen et al. claimed it was better to consider AoA ratings as a dependent variable (similar to word processing times) than as a predictor variable.

It is straightforward to test the claims of Zevin and Seidenberg (2002, 2004), Lété and Bonin (2013) and Baayen et al. (in press). Their basic message is that AoA ratings correlate with word processing times, not because they reflect the order in which words have been learned, but because they are contaminated by the perceived ease of the words. If a 'better' AoA measures is used, no effect of order of acquisition on word processing times will be left. A corollary of these claims is that the 'better' AoA measures must correlate more with the real order of acquisition than AoA ratings. So, all one has to do, is to demonstrate that the new, statistically derived measures are more related to objective, performance-based measures of AoA than the original ratings. Surprisingly, the only data we could find went in the opposite direction. Bonin et al. (2004) reported that, for the sample of 190 French words they tested, rated AoA correlated .68 with objective AoA (measured with picture naming like in Morrison et al., 1997), whereas frequency trajectory only correlated .32 with the criterion. One would have understood if this negative finding gave Bonin and colleagues pause for thought, but they cast it aside based on Zevin and Seidenberg's (2002, 2004) argument that the picture naming task cannot be used as an estimate of true AoA. If a more faithful criterion had been used of the order of acquisition, frequency trajectory would have correlated more with it than rated

AoA. In the present paper, we have a look at the various criterion data available in English to find out whether indeed the new, 'improved' AoA measures correlate more with faithful criteria of order of acquisition, as predicted by Zevin and Seidenberg (2002, 2004), Lété and Bonin (2013; Bonin et al., in press) and Baayen et al. (in press)..

Method

Validation data. The stimuli used as validation criterion consist of four tables with the ages at which words are supposed to be acquired / taught. The first table comes from the Morrison et al. (1997) study, in which children of various ages were asked to name 297 pictures of objects. As indicated above, this study has been criticized by Zevin and Seidenberg (2002, 2004) and Lété and Bonin (2013), because it confounds AoA with picture and/or name difficulty. It is also important to keep in mind that the Morrison et al. (1997) study happened in the UK, as word measures based on American participants explain less of variance for British individuals (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

The second table consists of the data from MarArthur-Bates Communicative Development Inventories (CDI). In this widely used test to detect delays in language development, parents of children from 8 months to 30 months are asked to indicate which words their child speaks. On the basis of the norming data, Goodman, Dale, and Li (2008) calculated the age of production for 562 words as the age (in months) at which 50% of the children were able to say the word.[2] This database is particularly interesting, because it deals with the first acquired words, at an age too young to have any explicit recollection of. Just like the Morrison et al. (1997) data, it deals with word production, but no longer on the basis of pictorial input. Łuniewska et al. (in press) showed that AoA ratings correlate around .50 with the CDI estimates in a variety of languages.

[2] With thanks to Philip S. Dale, who kindly provided the data.

The third database is the living word vocabulary (LWV) compiled by Edgar Dale and colleagues in the 1970s and 1980s, and published in Dale and O'Rourke (1981). The list includes more than 43,000 entries, separated on the basis of word meanings (i.e., there are different entries for the different meanings of homographs). For each item, the list provides a grade level (4, 6, 8, 10, 12, 13, or 16) at which the word is supposed to be known. The score was obtained by administering a three-choice test to pupils from schools in the Midwest of the United States. A grade was assigned as the level at which the percentage of correct responses exceeded 50% (when corrected for guessing). We used the lowest score as an estimate of AoA for homographs, as this is the first age participants learned the word. Biemiller and Slonim (2001) confirmed that words from LWV are learned in roughly the same order by children with different vocabulary sizes. Interestingly, Biemiller, Rosenstein, Sparks, Landauer, and Foltz (2014) in a recent study reported a correlation .58 between the LWV measure and AoA ratings based on a sample of 111 words.

The final dataset is a dataset very similar to LWV, provided by a website for American teachers (https://www.flocabulary.com/wordlists/; retrieved on December 4, 2015), which gives lists of words to be taught in various classes (going from Kindergarten to Grade 8). According to the website, these lists have been created by first compiling words from grade-appropriate novels and basal readers. The researchers then analyzed how often these vocabulary words appeared on US state tests. For each reading level, they looked at that level in the state tests and two grade levels above. So, the words taught in Grade 5 were defined as those words that are both found in 5th grade reading material and are most likely to appear on state tests in 5th, 6th and 7th grades. Information is available for 1,461 words. This makes it possible to see to what extent the LWV estimates still apply today. The correlation between the LWV grades and the Flocabulary grades is .69 for the 1,286 overlapping words.

The four datasets are summarized in Table 1. The table also shows the corresponding AoA ratings, obtained from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). As these authors already

noticed, people tend to overestimate the age at which they learned the first-acquired words. Very few participants give ratings below the age of 3, which corresponds to the time of infantile amnesia (Madsen & Kim, 2016). Interestingly, Łuniewska et al. (in press) observed that participants gave AoA estimates of almost half a year younger when they were asked "When do children learn this word?" instead of "When did you learn this word?". The correlation between both datasets (r = .93) was as high as could be expected based on the reliability of the ratings. So, participants seem to be particularly good at estimating the relative order of word acquisition, but not at giving the exact ages of acquisition.

- -

Insert Table 1 here

- -

Of further importance to notice is that the range of AoA values differs substantially between the four datasets. The LWV dataset not only is the largest, but also the one with the most variability in AoA values. In particular for the CDI dataset the range is very narrow, given that it deals with the very first words learned by children. All else equal, it is more difficult to find correlations in datasets with a small variability. This is known as the "restricted range" problem (Bland & Altman, 2011). As a matter of fact, given that all CDI words come from a very small range of words, acquired at an age of which the participants have no conscious recollection, one shouldn't be too surprised if no correlation were found between AoA ratings and CDI values, as argued by Baayen et al. (in press).

Predictor variables. As indicated above, the AoA ratings were obtained from Kuperman et al., (2012). They are based on Amazon Mechanical Turk users, who estimated the age in years at which they acquired the words. The measure is available for over 30,000 English words.

We used the SUBTLEX word frequencies, based on subtitles, as these have been shown to correlate best with human performance. We used the American frequencies (Brysbaert & New, 2009) for the American studies and the UK frequencies for the Morrison et al. dataset (Van Heuven et al., 2014). Brysbaert and Cortese (2010) showed that the effect of AoA is likely to be overestimated if a suboptimal word frequency measure is used. Word frequencies were expressed as Zipf values, which is a logarithmic scale (Van Heuven et al., 2014).

As the AoA effect is thought to be partly semantic in origin (Baayen et al., in press; Brysbaert & Ellis, in press), we also included two semantic variables. The first is word concreteness, based on the ratings collected by Brysbaert, Warriner, and Kuperman (2014) for over 40,000 English words. The second is the number of meanings (synsets) made available in Wordnet (https://wordnet.princeton.edu/). This dictionary lists for each word how many different meanings it has. For instance, the word "ball" can be used both as a noun and a verb. In the former part-of-speech, it has 12 different meanings, in the latter one, making a total of 13.

We also included word length, defined as the number of letters in the word.

Finally, the frequency trajectory measure was based on the TASA corpus, recommended by Zevin and Seidenberg (2002, 2004). This corpus contains word frequencies based on school books for different grades, going from grade 1 to 13+. In line with Zevin and Seidenberg, we compared the frequencies of the first three grades with those of the last three grades. As recommended by Bonin and colleagues, we calculated the frequency trajectory by first taking the logarithms of the frequencies and then transforming them to z-values for the low grades and the high grades separately. The frequency trajectory was defined as the difference between the z-value of the high

grades minus the z-value of the low grades. Bonin and colleagues further advised to calculate the cumulative frequency of a word as a better frequency measure, because it takes into account frequency of occurrence both at a young age and in adulthood. They defined cumulative frequency as the sum of the z-scores, which we did as well.

Results

Frequency trajectory and cumulative frequency. For the Morrison et al. (2007) database, we had data on all variables for 268 words. Table 2 shows the correlations between the variables. It clearly shows that the correlation between frequency trajectory and the criterion (the objective AoA estimates based on picture naming by children) is significantly lower than the correlation between rated AoA and the criterion: $r = .164$ vs. $.661$ ($z = -7.24$, $p < .001$), replicating Bonin et al. (2004) in English. At the same time, the table shows that the correlation between frequency trajectory and rated AoA remains significant ($r = .294$, $N = 268$, $p < .001$), so that frequency trajectory is likely to have a significant effect on AoA ratings even if it does not correlate with word processing times (Lété & Bonin, 2013). Cumulative frequency also correlates less with the criterion than rated AoA= $r = .489$ vs. $.661$ ($z = -2.99$, $p < .01$). When a multiple regression analysis is used to predict performance-based AoA on the basis of rated AoA, frequency trajectory, and cumulative frequency, rated AoA comes out as the best predictor ($t(264) = 9.31$, $p < .001$). Cumulative frequency is borderline significant ($t(264) = -2.06$, $p < .05$), and frequency trajectory is not involved at all ($t(264) = .13$). According to Bonin et al. (2004), this is acceptable as long as the objective AoA criterion is picture naming, but the situation should be different for other tasks. There, frequency trajectory should outperform rated AoA.

- -

Insert Table 2 here

- -

For the CDI dataset (first words produced by toddlers), we had all data for 513 words. Table 3 shows the results. Again, the correlation between frequency trajectory and the performance-based AoA criterion is substantially lower than that between rated AoA and the criterion: r = .279 vs. .458 (z = -3.38, p < .001). Frequency trajectory correlates significantly with rated AoA (r = .210, N = 531, p < .001). When a multiple regression analysis is used to predict the performance-based AoA criterion on the basis of rated AoA, frequency trajectory, and cumulative frequency, both AoA (t(509) = 11.90, p < .001) and cumulative frequency (t(509) = 4.95, p < .001) are significant, but frequency trajectory is not (t(509) = .17).

- -

Insert Table 3 here

- -

The LWV dataset (in which grade words should be taught) is the largest one with over 43,000 entries included. We had all data for 8,916 words. Table 4 shows the results. As before, the correlation between frequency trajectory and the performance-based AoA criterion is substantially lower than that between rated AoA and the criterion: r = .289 vs. .621 (z = -28.65, p < .001). Frequency trajectory correlates significantly with rated AoA (r = .501, N = 8,916, p < .001). In a multiple regression analysis, rated AoA correlates most with the criterion (t(8912) = 47.12, p < .001). Cumulative frequency reaches significance as well (t(8912) = -3.26, p .01), but frequency trajectory does not (t(8912) = -.05).

- -

Insert Table 4 here

- -

Finally, the Flocabulary database replicated the results with the LWV dataset. We had all data for 1,145 of the original 1,432 words. Table 5 shows the results. As in all previous analyses, the correlation between frequency trajectory and the performance-based AoA criterion is substantially lower than that between rated AoA and the criterion: r = .471 vs. .754 ($z = -11.25$, $p < .001$). Frequency trajectory correlates significantly with rated AoA (r = .453, N = 1,145, $p < .001$). When the performance-based AoA criterion is predicted on the basis of rated AoA, frequency trajectory, and cumulative frequency, the effects of all three variables are significant (AoA: $t(1141) = 16.60$; cumulative frequency: $t(1141) = -11.40$; frequency trajectory: $t(1141) = 11.58$). The performance-based AoA estimates were higher for words with high AoA ratings, infrequent words, and words that were more frequently used towards adults than towards children.

- -

Insert Table 5 here

- -

AoA ratings corrected for frequency. Baayen et al. (in press) propose to correct AoA ratings for frequency and other variables. Here we corrected them for frequency, word length, number of synsets, and concreteness (variables that are all significantly correlated with word frequency; see Tables 2-5).[3] We used two approaches. First, we took the residuals of a linear regression analysis including the confounded variables as predictors. Second, we ran a principal components analysis with orthogonal components on the predictors, as recommended by Baayen (2008, pp. 118-122). For all datasets, three principal components explained more than 5% of variance.

[3] The results are very similar when AoA ratings are corrected for word frequency only.

Table 6 shows the outcome of the analyses. In this table, the various correlations with the performance-based AoA criterion variables are listed. As in the analyses with frequency trajectory, the correlation between the criterion variable and rated AoA was higher than that between the criterion variable and the 'corrected' variables (all ps < .001) for three databases. The only exception was the CDI database (word knowledge in toddlers), where the corrected ratings correlated slightly better than the AoA ratings. The difference was not significant, though ($z = -0.71$, $p = .478$). AoA was extracted as the third principal component in the PC analysis.

- -

Insert Table 6 here

- -

The value of objective AoA estimates vs. AoA ratings to predict lexical decision times. A final set of analyses investigated whether performance-based AoA estimates are inferior to AoA ratings to predict lexical decision times. This prediction follows from Bonin and Baayen's claim that the high correlation between rated AoA and lexical decision time is due to differences in word frequency and word processing ease. To test the hypothesis, we used the standardized lexical decision times from the English Lexicon Project (Balota et al., 2007). Table 7 shows the outcome. The table illustrates that although rated AoA correlates more with lexical decision times than the performance-based estimates of AoA, this difference largely disappears when frequency is added as an additional predictor in the regression analysis. The latter is due to the fact that rated AoA correlates more with word frequency than performance-based AoA.

- -

Insert Table 7 here

- -

To test Bonin's claim that objective AoA would no longer influence lexical decision times, once cumulative frequency and frequency trajectory are taken into account, we ran multiple regression analyses with these three predictors on the ELP lexical decision tasks for each of the four stimulus sets. In each set, the effect of performance-based AoA remained a major contributor (Morrison et al.: $t(264) = 3.53$, $p < .001$); CDI: $t(507) = 3.97$, $p < .001$; LWV: $t(8827) = 23.33$, $p < .001$; Flocabulary: $t(1141) = 7.70$, $p < .001$)) In contrast, the effect of frequency trajectory was only significant for the largest LWV stimulus set (Morrison et al.: $t(264) = 1.38$; CDI: $t(507) = .13$; LWV: $t(8827) = 7.26$; Flocabulary: $t(1141) = -.76$).

Discussion

Authors have criticized AoA research because it is based on AoA ratings. Alternatives that have been proposed, are (1) objective AoA estimates based on the performance of children, (2) frequency trajectories based on the relative frequencies of words in young and older age, and (3) AoA ratings 'corrected' for confounding variables.

The present set of analyses confirms that performance-based AoA estimates do better than AoA ratings, because they are less correlated with word frequency, while at the same time being well correlated with word processing efficiency (as measured by lexical decision times; Table 7). These performance-based estimates consist of parental reports about when toddlers start to speak words, studies determining when preschool and primary school children can name pictures, studies establishing the age at which school children can select correct definitions of words, and studies estimating the age at which words are introduced in school curricula. There does not seem to be a difference between studies in which participants have to produce a name to a picture and studies with other dependent variables, contrary to what was claimed by Zevin and Seidenberg (2002, 2004) and Lété and Bonin (2013).

The present set of analyses could not find evidence for a superiority of frequency trajectory. As a matter of fact, it looks like one of the worst word characteristics ever introduced. In all datasets frequency trajectory correlated considerably less with the performance-based AoA criterion than rated AoA, as shown in Tables 2-5. The correlation was not completely absent (just like there remained a small correlation with rated AoA) but it was much reduced, arguably because the frequency trajectory variable contains a lot of noise. There are many reasons why frequency trajectory may not be a good estimate of AoA. The first is that the conversion from frequency of exposure to order of acquisition may not be straightforward. Goodman et al. (2008) give an interesting example. Although parents use lots of function words (prepositions, articles) when they talk to their toddlers, these words are not the first picked up by the children. As a matter of fact, Goodman et al. (2008) found a correlation between frequency of exposure and order of acquisition in toddlers only when they made a distinction between several word types (verbs, nouns, adjectives, interjections, function words, …). Second, given that children learn several new words per day, a few presentations may be enough for a word to be acquired. This is particularly true for words that interest children and that resemble existing concepts (e.g., unicorn), which may give rise to one-shot learning. Whereas the word frequency effect assumes that each encounter with a word has an equal impact, it looks like some words are easier to acquire than other (Brysbaert et al., 2016). Finally, there is the problem of how to compare word frequencies across ages. The relative numbers of words differ and the registers tapped into by the corpus are likely to differ as well. This clouds the definition of a frequency trajectory measure.

Also the analyses with cumulative frequency ran against the predictions made Zevin and Seidenberg and Bonin and colleagues. For a start, one prediction is that performance-based AoA measures (contrary to rated AoA measures) will no longer predict lexical decision times once cumulative frequency is controlled for. This is not what we observed in any of the datasets we examined. Another prediction is that rated AoA will correlate more with cumulative frequency than

with performance-based AoA. Again this is not the case in any of the analyses we did, as can be seen in Tables 2-5.

Correcting AoA ratings for frequency and other variables, as suggested by Baayen et al. (in press) hurts the correlation with the performance-based AoA validation criterion as well (Table 6). Very similar results are obtained when residual scores are used or values based on a principal components analysis (although in the latter case, the direction of the variables – positive or negative –becomes quite arbitrary, depending on the stimulus sample). It is true that AoA ratings have higher correlations with word frequency than performance-based AoA estimates, but the best way to take this into account is by entering both variables in a simultaneous regression analysis, rather than trying to correct one for the other (Wurm & Fasicaro, 2014). Because AoA and word frequency are intrinsically interrelated in real-life learning, there is no point in trying to obtain 'pure' measures of them, cleansed for the other variable. Indeed, the combined impact of AoA and word frequency is very similar for AoA ratings and objective AoA estimates (Table 7), indicating that a simultaneous regression analysis corrects quite well for the observation that AoA ratings correlate more with word frequency than performance-based estimates of AoA.

The fact that frequency trajectory and 'corrected' AoA ratings correlate less with the criterion variable than AoA ratings, means that one should be very cautious in interpreting null effects of the 'improved' variables. Given that they are less good estimates of the underlying construct, null effects become probable, particularly when the original effect is modest to start from. This is probably what happened in Lété and Bonin (2013). These authors reported a significant effect of frequency trajectory on AoA ratings but not on word processing efficiency. Given that the correlation between AoA and word processing efficiency is rather low to start from (Table 7), one should not be too surprised to observe a null effect when a mediocre predictor is used.

If authors feel unhappy about AoA ratings, it looks like there is only one valid alternative to follow. It consists of trying to collect more performance-based measures of AoA, based on studies that gauge word knowledge in children of various ages. Zevin and Seidenberg (2002 2004) and Lété and Bonin (2013) had a point when they said that it is better not to use picture naming in children as an estimate of the AoA effect in adult picture naming, because the task overlap by itself could account for any correlation observed. However, as the present study shows, the picture naming data do not diverge from those of the other tests used, and there are alternative formats to be tried out (based on teaching materials and multiple choice questions). An interesting research question in this respect will be to what extent the order of acquisition depends on the test used. Given that acquiring and being able to produce the full meaning of words is a protracted process, there arguably is not a single moment at which words are 'acquired'. It is well known in vocabulary research that scores on tests not only depend on the words used but also on the task to be performed (e.g., Pellicer-Sanchez, in press). Scores are lower when participants have to produce detailed definitions of the words or when the answer alternatives involve small changes in meaning (e.g., it is more difficult to give the right answer when the alternatives to choose from consist of 'big' and 'great' than when they are 'big' and 'green'). An interesting study in this respect will be to see to what extent the order of acquisition remains the same across tasks, even though the ages of 'acquisition' may depend on the format. The data of the present study make it quite likely that high correlations will be found in the order of acquisition between different test formats. Indeed, the pattern of results in the present study looked pretty stable, independent of how the performance-based measure of AoA had been operationalized.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press.

Baayen, R.H., Milin, P., & Ramscar, M. (in press). Frequency in lexical processing. *Aphasiology*.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39(3),* 445-459.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading, 18(2),* 130-154.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93(3),* 498-520.

Bland, J. M., & Altman, D. G. (2011). Correlation in restricted ranges of data. *BMJ, 342*:d556.

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and language, 50(4),* 456-476.

Bonin, P., Lété, B., Méot, A., Roux, S., & Ferrand, L. (in press). At what age did you learn "dog", "harp" and other words that you know? The issue of what adult age of acquisition (AoA) estimates really measure. In *Learning and Memory: Processes, Influences and Performance.* New York: Nova Science Publisher.

Bonin, P., Méot, A., Mermillod, M., Ferrand, L., & Barry, C. (2009). The effects of age of acquisition and frequency trajectory on object naming: Comments on Pérez (2007). *The Quarterly Journal of Experimental Psychology, 62(6),* 1132-1140.

Brysbaert, M., & Cortese, M. J. (2010). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology, 64(3),* 545-559.

Brysbaert, M. & Ellis, A.W. (in press). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41(4),* 977-990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance ;42(3*),441-458.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46(3),* 904-911.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture naming latency. *The Quarterly Journal of Experimental Psychology, 25(1),* 85-95.

Catling, J., Dent, K., Preece, E., & Johnston, R. (2013). Age-of-acquisition effects in novel picture naming: A laboratory analogue. *The Quarterly Journal of Experimental Psychology, 66(9),* 1756-1763.

Dale, E., & O'Rourke, J. (1981). *The Living Word Vocabulary, the Words We Know: A National Vocabulary Inventory*. Chicago, IL: World book.

De Moor, W., Ghyselinck, M. & Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor & Brysbaert. *Psychologica Belgica, 40,* 99-114.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(5),* 1103-1123.

Gilhooly, K. J., & Gilhooly, M. L. M. (1980). The validity of age‐of‐acquisition ratings. *British Journal of Psychology, 71(1),* 105-110.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35(3),* 515-531.

Joseph, H. S., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133(1),* 238-248.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44(4),* 978-990.

Lété, B., & Bonin, P. (2013). Does frequency trajectory influence word identification? A cross-task comparison. *The Quarterly Journal of Experimental Psychology, 66(5),* 973-1000.

Łuniewska, M., Haman, E., Armon-Lotem, E., Etenkowski, B., Southwood, F., ... & Unal-Logacev, O.. (2014). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*.

Madsen, H. B., & Kim, J. H. (2016). Ontogeny of memory: An update on 40 years of work on infantile amnesia. *Behavioural Brain Research, 298,* 4-14.

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language, 63(4),* 506-525.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology: Section A, 50(3),* 528-559.

Pellicer-Sanchez, A. (In press). Incidental vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*. DOI: http://dx.doi.org/10.1017/S0272263115000224.

Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *The Quarterly Journal of Experimental Psychology, 61(3),* 361-371.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67(8),* 1176-1190.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72(1),* 37-48.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and language, 47(1),* 1-29.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition, 32(1),* 31-38.

Table 1: Datasets used for the validation. N gives the number of stimuli for which AoA ratings and frequency trajectories are available. The age range indicates the age (in years) at which the words were supposed to be acquired. For LWV and Flocabulary, the exact minimum and maximum ages could not be ascertained, as some words were known to more than half of the youngest pupils tested or less than half of the oldest students tests. The table also shows the corresponding AoA ratings from Kuperman et al. (2012), the Zipf frequency, and frequency trajectory data, and the cumulative frequency.

	$N_{stimuli}$	Age range (years)	AoA_{rating} (SD)	Range ratings	Frequency Zipf (SD)	FreqTraj (SD)	Cumfreq (SD)
Morrison et al.	286	1.8 – 11.7	5.6 (1.6)	2.9–11.9	4.22 (.58)	-1.08 (0.68)	1.60 (1.77)
CDI	531	1 – 2.5	4.4 (0.9)	1.6– 8.5	4.91 (.93)	-0.98 (0.67)	3.34 (2.46)
LWV	20,004	<8 - >16	10.4 (3.0)	1.8–21.0	3.66 (.82)	0.09 (1.01)	.30 (1.88)
Flocabulary	1,432	<6 - >14	9.2 (2.4)	3.8–15.8	3.77 (.64)	0.41 (0.96)	.56 (1.59)

Table2: Pearson correlation coefficients between the objective AoA estimates published by Morrison et al. (1997) and the predictor variables (N = 268, values are significant at .05 level when r > .120; the results are very similar with Spearman rank correlations)

	AoA_{rating}	$Freq_{Zipf}$	Length	Synsets	Concrete	FreqTraj	Cumfreq
AoA_{crit}	0.661	-0.547	0.233	-0.250	-0.188	.164	-.470
Aoa_{rating}		-0.587	0.422	-0.307	-0.249	.294	-.589
$Freq_{Zipf}$			-0.472	0.572	0.072	.132	.859
Length				-0.438	-0.003	.182	-.483
Synsets					-0.200	.153	.534
Concrete						-.116	.000
FreqTraj							.096

Table 3: Pearson correlation coefficients between the CDI objective AoA estimates published by Goodman et al. (2008) and the predictor variables (N = 513, values are significant at .05 level when r > .085; the results are very similar with Spearman rank correlations)

	AoA_{rating}	$Freq_{Zipf}$	Length	Synsets	Concrete	FreqTraj	Cumfreq
AoA_{crit}	0.458	0.109	0.031	0.258	-0.101	.279	.133
Aoa_{rating}		-0.280	0.261	-0.206	-0.125	.210	-.263
$Freq_{Zipf}$			-0.508	0.642	0.048	.415	.904
Length				-0.228	0.063	-.158	-.510
Synsets					0.225	.280	.610
Concrete						-.052	-.005
FreqTraj							.551

Table 4: Pearson correlation coefficients between the objective LWV AoA estimates published by Dale and O'Rourke (1981) and the predictor variables (N = 8,916, values are significant at .05 level when r > .021; the results are very similar with Spearman rank correlations)

	AoA_{rating}	$Freq_{Zipf}$	Length	Synsets	Concrete	FreqTraj	Cumfreq
AoA_{crit}	0.621	-0.430	0.239	-0.203	-0.311	.289	-.406
Aoa_{rating}		-0.621	0.434	-0.290	-0.373	.501	-.609
$Freq_{Zipf}$			-0.410	0.440	0.040	-.147	.832
Length				-0.296	-0.316	.412	-.357
Synsets					0.054	-.037	.463
Concrete						-.485	.052
FreqTraj							-.060

Table 5: Pearson correlation coefficients between the objective AoA estimates published on https://www.flocabulary.com/wordlists/ and the predictor variables (N = 1,145, values are significant at .05 level when r > .058; the results are very similar with Spearman rank correlations)

	AoA_{rating}	$Freq_{Zipf}$	Length	Synsets	Concrete	FreqTraj	Cumfreq
AoA_{crit}	0.754	-0.529	0.367	-0.285	-0.443	.471	-.587
Aoa_{rating}		-0.597	0.356	-0.334	-0.479	.453	-.635
$Freq_{Zipf}$			-0.221	0.442	0.092	-.061	.810
Length				-0.234	-0.338	.435	-.253
Synsets					0.178	-.039	.465
Concrete						-.470	.232
FreqTraj							-.046

Table 6: Pearson correlations between the various AoA estimates and the criterion variables. The number of stimuli in some cases are higher than in the previous tables, because this time the words not present in the frequency trajectory table did not have to be excluded.

	N	AoA_{rating}	AoA_{Res}	PCA1	PCA2	PCA3
Morrison et al.	286	0.640	0.345	-0.282	0.120	-0.345
CDI	531	0.453	0.487	-0.198	0.231	0.497
LWV	20,004	0.727	0.420	0.322	-0.069	-0.420
Flocabulary	1,432	0.778	0.384	0.368	-0.252	0.384

Table 7: Percentages of variance in ELP lexical decision times accounted for AoA ratings and objective estimates of AoA, both when considered in isolation and together with word frequency. The number of stimuli depends on the number of data present both in the original databases and in ELP. All models are significant at the $p < .01$ level.

	N	AoA_{rating}	AoA_{obj}	AoA_{rat}+freq	AoA_{obj}+freq
Morrison et al.	283	0.368	0.190	0.502	0.443
CDI	513	0.022	0.015	0.084	0.103
LWV	17,309	0.444	0.292	0.549	0.517
Flocabulary	1,428	0.389	0.317	0.506	0.493